# Una herramienta basada en teoría de grafos y teoría de la información para la caracterización del entorno de interacción de un conjunto de proteínas.

R. Massanet Vila[1,2,3], J.J. Gallardo Chacón[3], P. Caminal Magrans[1,2,3], A. Perera Lluna[1,2,3]

[1]Dept. ESAII, Universitat Politècnica de Catalunya (UPC), Barcelona, España;
{raimon.massanet, pere.caminal,alexandre.perera}@upc.edu
[2]Centre de Recerca en Enginyeria Biomèdica (CREB), Barcelona, España;
[3]CIBER de Bioingeniería, Biomateriales y Nanomedicina (CIBER-BBN), España;
{joan.josep.gallardo}@upc.edu

## Resumen

*En los últimos años se han acumulado grandes cantidades de información genética, proteómica y metabólica en bases de datos públicas y privadas. Se ha dedicado un gran esfuerzo al desarrollo de herramientas capaces de explotar, estructurar y organizar esta información para poder sacar partido de ella. Sin embargo, a día de hoy no existe ninguna aplicación que caracterice el entorno en el que una proteína o conjunto de proteínas actúa. En este trabajo se ha desarrollado un paquete de software que consulta, de forma masiva, bases de datos públicas, como* Gene Ontology Annotation *(GOA) y* Human Proteome Resource Database *(HPRD) y presenta la información en ellas contenida de forma que pueda ser fácil y rápidamente interpretada por personas pertenecientes al entorno clínico y no familiarizadas con el uso de bases de datos. Para ello presenta al usuario un conjunto de grupos de proteínas que forman parte del entorno de interacción local del conjunto inicial de proteínas y que forman subredes de alta conectividad. Además describe con etiquetas semánticas los aspectos particulares de cada grupo.*

## 1. Introducción

La mayor parte de los procesos biológicos de la célula necesitan actividades combinadas y sincronizadas de conjuntos de proteínas que forman cadenas metabólicas de señalización y regulación [1]. Los datos experimentales sobre proteínas y su papel en el funcionamiento de los organismos son recogidos y organizados en grandes bases de datos que son de gran utilidad para la extracción de relaciones biológicas [2]. Sin embargo, la información sobre sistemas biológicos debe ser organizada y priorizada para atender a las características específicas que cada área de investigación requiere. Multitud de herramientas bioinformáticas aparecen y son evaluadas día a día [3, 4], con la intención de resolver algunos de estos problemas prácticos. Además, ontologías, como la *Gene Ontology* (GO) [5], han sido definidas para intentar estandarizar la forma en que los diferentes grupos de investigación alrededor del mundo anotan genes y proteínas con etiquetas semánticas que definen aspectos importantes como su función biológica o el componente celular en el que actúan, por ejemplo. Estas ontologías permiten, así mismo, que dicha información semántica sea minada, procesada e incluso generada por programas informáticos.

Entender los procesos celulares y las mutaciones genéticas que les afectan es un objetivo de vital importancia en esas disciplinas. La gran cantidad de datos acumulados en los últimos años podría ser crucial en el futuro desarrollo de la genética, la proteómica y la metabolómica. Sin embargo, consultar y explotar grandes cantidades de información puede ser todo un reto. Las técnicas de *clustering*, o agrupamiento, son bien conocidas y ampliamente aceptadas por la comunidad científica para el análisis exploratorio de grandes cantidades de datos. Partir un gran conjunto de datos en grupos de menor tamaño y mayor coherencia ayuda a entender mejor la estructura subyacente en los datos. Recientemente, la técnica de *spectral clustering* [6, 7], o agrupamiento espectral, ha ganado una gran popularidad. Se ha demostrado que el agrupamiento espectral resuelve el problema de partir un grafo de forma óptima desde diferentes puntos de vista al mismo tiempo: desde el punto de vista del mínimo corte del grafo (*Min Graph Cut*), desde el enfoque de caminos aleatorios (*Random Walk*), y usando teoría de la perturbación [8].

Combinando información de interacciones entre proteínas, agrupamiento espectral e información semántica es posible enriquecer las redes de interacción entre proteínas (RIP) [9] y extraer información valiosa sobre los diferentes grupos hallados para explicar las relaciones entre proteínas. Una vez organizada, esta gran cantidad de información podría ser de gran utilidad para el estudio de problemas macroscópicos tales como patologías o modelos fisiológicos.

En este trabajo se propone una metodología para ex-

traer, de forma automática, información de bases de datos públicas, y organizarla para que se pueda extraer de ella un conocimiento útil sobre el entorno local de interacción de una proteína o grupo de proteínas. Esto se consigue mediante la división de dicho entorno en diferentes subredes de interacción de alta densidad y encontrando las etiquetas semánticas de cada subred que mejor las caracterizan. Este proceso podría ser de gran utilidad para investigadores dedicados al estudio de un conjunto de proteínas pues les permitiría obtener, de forma rápida y automática, una descripción de los procesos que acontecen en las inmediaciones de la proteína, o grupo de proteínas, en cuestión. Además, esta descripción es presentada en un lenguaje casi natural y usando un estándar mundial de anotación. Se ha desarrollado un paquete software en el lenguaje de programación estadística R [10] que implementa la metodología. Este software caracteriza los procesos en los que un conjunto de proteínas y sus vecinos más inmediatos participan.

## 2. Materiales y métodos

### 2.1. Metodología

El proceso desarrollado para la caracterización del entorno de interacción del conjunto de proteínas de entrada se basa en el agrupamiento espectral y la caracterización semántica de los grupos obtenidos. Para ello se construye una red de interacciones formada por las proteínas que rodean al conjunto de entrada. A continuación se aplica la técnica de agrupamiento espectral para hallar grupos con una interconexión interna elevada. Finalmente se caracteriza cada uno de los grupos aplicando un proceso de enriquecimiento semántico. Este proceso se ha desarrollado en cuatro pasos, detallados a continuación. A lo largo del resto de este texto, el término *interacción* se refiere a interacción física (acoplamiento molecular) entre dos proteínas. $S$ representa el conjunto de proteínas de entrada, para el cual se quiere obtener una descripción del entorno de interacción.

#### *2.1.1. Minado de interacciones*

En el primer paso el algoritmo realiza un minado de la base de datos *Human Proteome Resource Database* (HPRD) [11] con el objetivo de construir un entorno de interacción local para el conjunto $S$. Este entorno contiene todas aquellas proteínas, contenidas en la base de datos, que se encuentran a menos de $n$ interacciones de distancia de cualquiera de las proteínas del conjunto $S$. La variable $n$ fue definida como parámetro para permitir al usuario el control de la cantidad de información del entorno que se quiere incorporar al estudio. La información obtenida de la base de datos HPRD es transformada a una estructura de grafo no dirigido, en el que los nodos representan proteínas y las artistas representan interacciones entre proteínas. A este grafo se le denomina *entorno de interacción local* del conjunto de proteínas $S$ y se denota $G_S$ [12].

#### *2.1.2. Agrupamiento*

En la segunda fase se aplica la técnica de agrupamiento espectral para partir el grafo $G_S$ obtenido en la fase anterior. El agrupamiento espectral es una técnica que se basa en la diagonalización de la matriz laplaciana del grafo a partir.

Para la creación de la matriz laplaciana primero se definen la matriz de grados $D$ y la matriz de pesos $W$. La matriz $D$ es una matriz cuya diagonal contiene los grados de los nodos del grafo. Es decir, $d_{ii}$ es el grado del nodo $i$ y $d_{ij} = 0$ para todo $i \neq j$. La matriz $W$ es una matriz de pesos entre los nodos del grafo. En este caso los pesos representan similitudes entre las proteínas. Por tanto, $w_{ij}$ es una medida de similitud entre los nodos $i$ y $j$. Gran variedad de medidas pueden ser definidas a efectos de construir la matriz $W$. En este trabajo se ha optado por usar la matriz de adyacencias del grafo como matriz de pesos:

$$w_{ij} = \begin{cases} 1 & \text{si las proteínas } i \text{ y } j \text{ interactúan} \\ 0 & \text{en caso contrario} \end{cases} \quad (1)$$

Sin embargo, se podrían utilizar otras medidas para definir la similitud entre dos proteínas. Por ejemplo, se podrían usar el número de publicaciones en las que aparecen las dos proteínas, una medida de similitud semántica basada en Gene Ontology [13] y otras.

Una vez definidas las matrices $D$ y $W$ se puede definir la matriz laplaciana. Diferentes autores han propuesto diferentes matrices laplacianas en la literatura. En este trabajo se ha seguido el texto de von Ulrike [8] y se han implementado tres matrices laplacianas diferentes.

La matriz laplaciana no normalizada:

$$L_u = D - W \quad (2)$$

La matriz laplaciana normalizada simétrica:

$$L_{sym} = I - D^{-1/2}\, W\, D^{-1/2} \quad (3)$$

Y la matriz laplaciana normalizada de caminos aleatorios (*random-walk*):

$$L_{rw} = I - D^{-1}\, W \quad (4)$$

La matriz laplaciana a usar para la división del grafo puede ser escogida por el usuario, aunque por defecto se usa la última por ser la que mejores resultados parece obtener.

Una vez construido el modelo espectral, el usuario debe decidir en cuantos grupos se quiere partir el entorno $G_S$. Esta suele ser una decisión crítica y difícil en todo proceso de agrupamiento. Sin embargo, el modelo espectral permite tomar esta decisión en base a unos fundamentos teóricos sólidos. La matriz laplaciana

tendrá tantos valores propios con valor 0 como componentes conexas tenga el grafo $G_S$. Además, valores propios próximos a 0 indicarán componentes del grafo con una alta densidad interna de aristas y una baja conexión con el resto del grafo. Esto permite tomar la decisión del número de grupos a obtener en base a la estructura del grafo. Para ello se presenta al usuario una gráfica con los valores propios de la matriz laplaciana. A continuación el usuario indica el número de grupos $k$ en los que desea partir los datos, que deberían ser tantos como valores propios próximos a 0. Este valor es un parámetro del modelo y es, en cierto modo, un factor de escala. A menor $k$ se obtendrán pocos grupos de gran tamaño, situación idónea para realizar un estudio a gran escala. A medida que $k$ aumenta se obtiene un mayor número de grupos de menor tamaño, lo que permite realizar un estudio a menor escala. Finalmente, se aplica un algoritmo de agrupamiento clásico, como *k-means*, para partir la matriz formada por los vectores propios de la matriz laplaciana en $k$ grupos.

#### *2.1.3. Minado semántico*

El tercer paso es minar la base de datos de Gene Ontology Annotation (GOA) [14] para enriquecer la red de interacciones, dividida en el paso anterior, con información semántica. Para cada proteína de $G_S$ se obtiene el correspondiente conjunto de anotaciones semánticas. A continuación se obtiene, para cada partición, la particular distribución de cada término semántico. Es decir, para cada término semántico se contabiliza el número de proteínas del grupo que están anotadas a ese término.

En esta fase se ofrece al usuario la posibilidad de filtrar las anotaciones semánticas que quiere utilizar para enriquecer la red de interacciones. Este filtrado se realiza en base al código de evidencia (*evidence code*) de las anotaciones. Este código indica el tipo de evidencia científica en base a la cual se ha introducido la anotación. Por ejemplo, es práctica habitual en la literatura científica excluir las anotaciones semánticas con código de evidencia IEA. Este tipo de anotaciones son inferidas electrónicamente por programas informáticos en base a homología, similitud estructural, etc, y muchos autores las descartan por no ser fruto de experimentos bioquímicos.

#### *2.1.4. Estadística*

Finalmente, se realizan pruebas estadísticas para encontrar aquellas anotaciones semánticas que mejor describen cada grupo de proteínas. A tal efecto, se compara la distribución de las anotaciones de cada grupo con una distribución nula utilizando una prueba de Mann-Whitney [15]. La distribución nula es generada para cada anotación $a$ de cada partición $p$ como la distribución de $a$ en 50 muestras de proteínas, del mismo tamaño muestral que $p$, tomadas aleatoriamente entre todas las que forman la red y no están en la partición $p$. Los $p$-valores fueron calculados empíricamente modelando la función de la distribución nula usando métodos basados en *kernels* [16]. Este proceso devuelve un nivel de significación empírica que cuantifica las diferencias entre la distribución de la anotación dentro y fuera de la partición. Seleccionando las anotaciones con mayor significación estadística el usuario obtiene una descripción semántica de las características particulares de cada grupo $p$ en oposición a las características de todo el entorno de interacción $G_S$. Este proceso permite realizar una descripción semántica modular del entorno de interacción del conjunto de proteínas inicial $S$.

### 2.2. Aplicación práctica

Con el objetivo de evaluar el comportamiento de la aplicación, así como mostrar su funcionamiento, se propuso un caso de aplicación práctica por parte de un experto en proteómica. La metodología fue aplicada al estudio del entorno de interacción de la proteína humana HSP27 (*small heat shock protein*). El parámetro $n$ se fijó a 3 porque este valor produjo una red con un alto compromiso entre cantidad de información elevada y manejabilidad de los datos. Se uso la matriz laplaciana de caminos aleatorios (*random-walk*) para construir el modelo espectral. La red de interacciones fue dividida en 8 grupos después de examinar la función de los valores propios del modelo espectral porque este valor definía un salto importante en su derivada. Tras partir y enriquecer semánticamente la red, se realizaron las pruebas estadísticas y se seleccionaron las 10 anotaciones con mayor significación para cada grupo. Los resultados de este caso práctico se presentan en la siguiente sección 3.

## 3. Resultados

El entorno de interacción local que se obtuvo para la proteína HSP27 consistió en una red de 601 proteínas y 2840 interacciones. La Figura 1 muestra la distribución de los grados de los nodos de la red, que se caracteriza por pequeño número de nodos con alta conectividad y un gran número de nodos de baja conectividad. 7 de las 8 particiones tuvieron al menos una etiqueta estadísticamente significativa que permitió realizar una interpretación semántica del módulo. La partición restante contenía solamente una anotación semántica que no resultó ser estadísticamente significativa. La Figura 2 muestra el número de proteínas en cada grupo. Los grupos 3, 7 y 8 contienen la mayor parte de las proteínas del entorno de interacción.

La proteína inicial, HSP27, fue asignada por la técnica de agrupamiento espectral al grupo 3, el cual contenía el mayor número de proteínas. Esta gran red de alta densidad de interacciones es típica de las redes *libres de escala* [17] e indica que la proteína HSP27 puede tener un rol importante en un conjunto de fun-

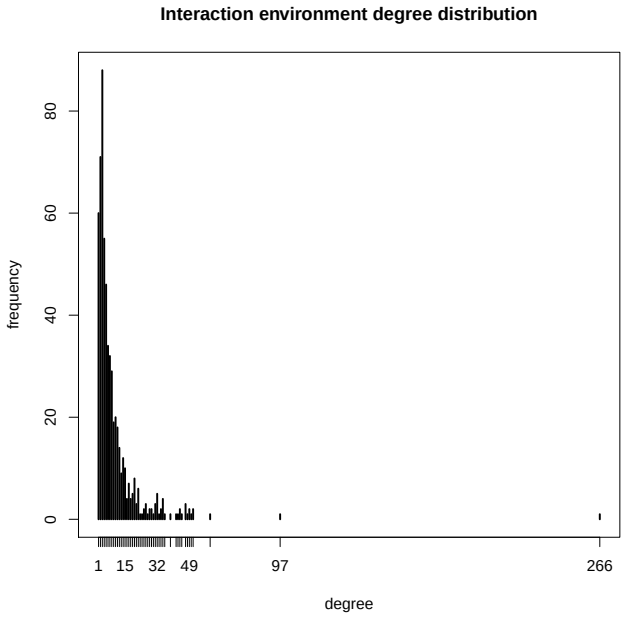


Figura 1: *Distribución de los grados de de la red de interacción local para la proteína HSP27 con un máximo de* 3 *niveles de interacción.*

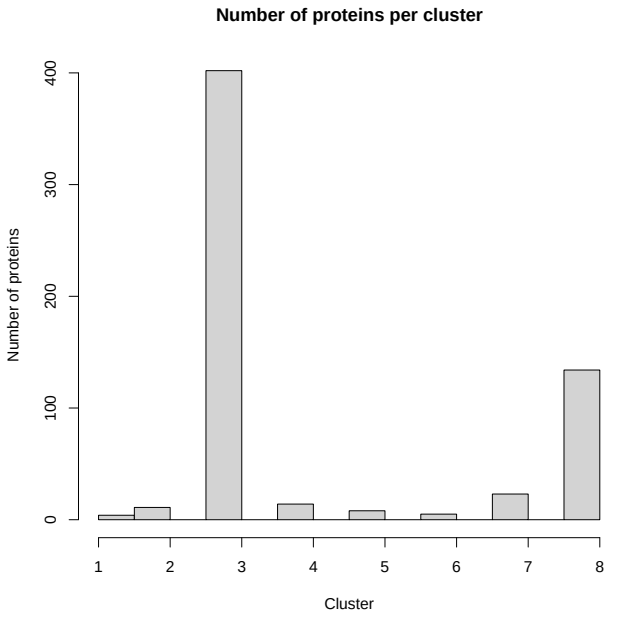


Figura 2: *Número de proteínas en cada grupo después de aplicar la técnica de agrupamiento espectral.*

ciones biológicas muy interconectadas. Este grupo fue caracterizado semánticamente por sus anotaciones estadísticamente significativas, las 10 primeras de las cuales pueden consultarse en la Tabla 1. En este proceso se utilizaron solamente etiquetas semánticas relativas a proceso biológico (BP) y se excluyeron las anotaciones inferidas electrónicamente (IEA). La columna *Dirección* de las Tablas 1, 2, 3. indica si la etiqueta tiene una frecuencia en el grupo mayor o menor que en la distribución nula. Los resultados sugieren que este grupo de proteínas tiene un papel fundamental en la regulación de varios procesos, incluyendo la producción de factor-beta2, la biosíntesis del óxido nítrico y la apoptosis.

Otras subredes de tamaño considerable y alta coherencia fueron halladas en los grupos 8 y 7. Este resultado sugiere una relación indirecta de la proteína HSP27 con otros procesos biológicos. El grupo 8 se diferencia del resto por las etiquetas halladas en la Tabla 3, que sugieren que consta de proteínas involucradas en respuesta a estímulos, virus y corrección de errores de transcripción. Por otra parte, el grupo 7 se describe en la Tabla 2 como un grupo de proteínas altamente involucrado en transporte de proteínas y movimiento celular.

Por razones de espacio, solamente se muestran en este texto las 10 etiquetas más significativas de cada grupo. Sin embargo, el usuario tiene la oportunidad de explorarlas todas y filtrarlas según un umbral de significación estadística.

| | Etiqueta | $p$-valor | Dirección |
|---|---|---|---|
| 1 | activation of JNK activity | 4.23e-19 | high |
| 2 | DNA ligation | 5.65e-19 | high |
| 3 | regulation of transforming growth factor-beta2 production | 5.65e-19 | high |
| 4 | positive regulation of nitric oxide biosynthetic process | 6.52e-19 | high |
| 5 | negative regulation of protein kinase activity | 1.27e-18 | high |
| 6 | nitric oxide biosynthetic process | 1.28e-18 | high |
| 7 | positive regulation of protein binding | 1.28e-18 | high |
| 8 | positive regulation of transcription | 1.35e-18 | high |
| 9 | base-excision repair | 1.35e-18 | high |
| 10 | induction of apoptosis | 1.47e-18 | high |

Cuadro 1: *Descripción semántica del grupo* 3 *con anotaciones de proceso biológico.*

La Tabla 4 muestra el número de etiquetas significativas para cada grupo. En ella puede observarse que todos los grupos tienen una proporción importante de etiquetas significativas. Este resultado indica que el agrupamiento espectral obtenido solamente a partir

| | Etiqueta | $p$-valor | Dirección |
|---|---|---|---|
| 1 | actin filament bundle formation | 1.34e-18 | high |
| 2 | actin cytoskeleton organization and biogenesis | 4.96e-18 | high |
| 3 | cell motility | 1.26e-17 | high |
| 4 | ovarian follicle development | 9.74e-03 | high |
| 5 | oocyte maturation | 9.74e-03 | high |
| 6 | diacylglycerol biosynthetic process | 9.74e-03 | high |
| 7 | calcium ion-dependent exocytosis | 9.74e-03 | high |
| 8 | actin filament polymerization | 9.74e-03 | high |
| 9 | actin filament-based movement | 9.74e-03 | high |
| 10 | actomyosin structure organization and biogenesis | 9.74e-03 | high |

Cuadro 2: *Descripción semántica del grupo* 7 *con anotaciones de proceso biológico.*

| | Etiqueta | $p$-valor | Dirección |
|---|---|---|---|
| 1 | postreplication repair | 8.15e-18 | high |
| 2 | RNA processing | 7.36e-03 | high |
| 3 | nuclear mRNA splicing, via spliceosome | 7.54e-03 | high |
| 4 | alcohol metabolic process | 9.74e-03 | high |
| 5 | inosine catabolic process | 9.74e-03 | high |
| 6 | unknown GO label | 9.74e-03 | high |
| 7 | segment specification | 9.74e-03 | high |
| 8 | response to biotic stimulus | 9.74e-03 | high |
| 9 | response to virus | 1.58e-02 | high |
| 10 | response to unfolded protein | 1.64e-02 | high |

Cuadro 3: *Descripción semántica del grupo* 8 *con anotaciones de proceso biológico.*

| | Etiquetas | Etiquetas significativas |
|---|---|---|
| 1 | 5 | 1 |
| 2 | 10 | 4 |
| 3 | 666 | 337 |
| 4 | 20 | 9 |
| 5 | 16 | 8 |
| 6 | 1 | 0 |
| 7 | 59 | 31 |
| 8 | 131 | 33 |

Cuadro 4: *Número de etiquetas semánticas en cada grupo.*

de la estructura de la red de interacciones define grupos con una también elevada coherencia semántica.

## 4. Conclusiones

En este trabajo se ha desarrollado un método para caracterizar de forma rápida y eficiente el entorno de interacción local de un conjunto de proteínas. Se ha presentado un caso práctico de aplicación para demostrar el funcionamiento del software desarrollado. Los resultados muestran que la partición obtenida define grupos de proteínas con una elevada coherencia semántica y etiquetas llenas de significado. El método realiza un minado automático y masivo de bases de datos públicas y sintetiza para el usuario toda la información en una descripción que utiliza un lenguaje casi natural, utilizando etiquetas semánticas ampliamente aceptadas.

## 5. AGRADECIMIENTOS

Los autores agradecen el apoyo recibido por parte del Ministerio de Educación y Ciencia a través del programa Ramón y Cajal y TEC2007-63637/TCM así como del Instituto de Salud Carlos III a través de la iniciativa CIBER-BBN en Bioingeniería , biomateriales y nanomedicina.

## Referencias

[1] M. Monti, M. Cozzolino, F. Cozzolino, G. Vitiello, R. Tedesco, A. Flagiello, and P. Pucci, "Puzzle of protein complexes in vivo: a present and future challenge for functional proteomics." *Expert Rev Proteomics*, vol. 6, no. 2, pp. 159–169, Apr 2009.

[2] M. Krallinger, F. Leitner, and A. Valencia, "Analysis of biological processes and diseases using text mining approaches." *Methods Mol Biol*, vol. 593, pp. 341–382, 2010.

[3] A. A. Terentiev, N. T. Moldogazieva, and K. V. Shaitan, "Dynamic proteomics in modeling of the living cell. protein-protein interactions." *Biochemistry (Mosc)*, vol. 74, no. 13, pp. 1586–1607, Dec 2009.

[4] K. Raman, "Construction and analysis of protein-protein interaction networks." *Autom Exp*, vol. 2, no. 1, p. 2, 2010.

[5] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock, "Gene ontology: tool for the unification of biology. the gene ontology consortium," *Nature genetics*, vol. 25, no. 1, pp. 25–29, May 2000.

[6] W. E. Donath and A. J. Hoffman, "Lower bounds for the partitioning of graphs," *IBM J.Res.Dev.*, vol. 17, no. 5, pp. 420–425, 1973.

[7] M. Fiedler, "Algebraic connectivity of graphs,"

*Czechoslovak Mathematical Journal*, vol. 23, no. 2, pp. 298–305, 1973.

[8] U. von Luxburg, "A tutorial on spectral clustering," *Statistics and Computing*, vol. 17, no. 4, pp. 395–416, December 2007.

[9] M. Deng, Z. Tu, F. Sun, and T. Chen, "Mapping gene ontology to proteins based on protein-protein interaction data." *Bioinformatics*, vol. 20, no. 6, pp. 895–902, Apr 2004.

[10] R Development Core Team, "R: A language and environment for statistical computing," 2009. [Online]. Available: http://www.R-project.org

[11] T. S. K. Prasad, R. Goel, K. Kandasamy, S. Keerthikumar, S. Kumar, S. Mathivanan, D. Telikicherla, R. Raju, B. Shafreen, A. Venugopal, L. Balakrishnan, A. Marimuthu, S. Banerjee, D. S. Somanathan, A. Sebastian, S. Rani, S. Ray, C. J. H. Kishore, S. Kanth, M. Ahmed, M. K. Kashyap, R. Mohmood, Y. L. Ramachandra, V. Krishna, B. A. Rahiman, S. Mohan, P. Ranganathan, S. Ramabadran, R. Chaerkady, and A. Pandey, "Human protein reference database–2009 update," *Nucleic acids research*, vol. 37, no. suppl_1, pp. D767–772, January 1 2009.

[12] R. Massanet-Vila, J. J. Gallardo-Chacón, P. Caminal, and A. Perera, "Búsqueda de genes candidatos mediante redes de interacciones entre proteínas y medidas de similitud semántica basadas en gene ontology: Aplicación al síndrome de brugada," in *XXVII Congreso Anual de la Sociedad Española de Ingeniería Biomédica, CASEIB2009*, 2009.

[13] C. Pesquita, D. Faria, H. Bastos, A. E. Ferreira, A. O. Falcao, and F. M. Couto, "Metrics for go based protein semantic similarity: a systematic evaluation," *BMC Bioinformatics*, vol. 9, p. S4, 2008.

[14] D. Barrell, E. Dimmer, R. P. Huntley, D. Binns, C. O'Donovan, and R. Apweiler, "The goa database in 2009–an integrated gene ontology annotation resource," *Nucl.Acids Res.*, p. gkn803, October 2008.

[15] H. B. Mann and D. R. Whitney, "On a test of whether one of two random variables is stochastically larger than the other," *The Annals of Mathematical Statistics*, vol. 18, no. 1, pp. 50–60, Mar. 1947.

[16] S. J. Sheather and M. C. Jones, "A reliable data-based bandwidth selection method for kernel density estimation," *Journal of the Royal Statistical Society.Series B (Methodological)*, vol. 53, no. 3, pp. 683–690, 1991. [Online]. Available: http://www.jstor.org/stable/2345597

[17] R. Albert, "Scale-free networks in cell biology," *Journal of cell science*, vol. 118, no. 21, pp. 4947–4957, November 1 2005.